28th IOPS Winter Conference
13-14 December 2018

Nederlands Watermuseum

Zijpendaalseweg 26-28, 6814 CL Arnhem

[ΨΣ] interuniversity graduate school of iops Psychometrics and Sociometrics

CITO 50 jaar

Prior to the conference – Thursday December 13th

10.30 – 12.00 **IOPS Board meeting** (Archicolourzaal)

11.30 – 12.00 **IOPS PhD student meeting** (Watercinema)

12.00 – 13.00 Lunch and registration (Grand Café Aan de Beek)

Program Thursday December 13th (Watercinema)

13.00 – 13.05 **Official opening by Cor Sluijter**
Head of Psychometric Research department, Cito

13.05 – 13:30 **Presentation** *Alexandra de Raadt, University of Groningen*
A comparison of agreement coefficients for categorical and interval scales
Discussant: Dylan Molenaar

13.30 – 13.55 **Presentation** *Nitin Bhushan, University of Groningen*
Comparing Constraint-based Causal Discovery algorithms in scenarios typical to psychology
Discussant: Robbie van Aert

13.55 – 14.20 **Presentation** *Sara van Erp, Tilburg University*
Shrinkage priors for Bayesian measurement invariance:
Practical and robust approaches for modeling and detecting non-invariance
Discussant: Leonie Vogelsmeier

14.20 – 14.45 **Presentation** *Konrad Klotzke, University of Twente*
Bayesian Covariance Structure Modelling of Responses and Process Data
Discussant: Herbert Hoijtink

14.45 – 15.15 Break (Watercinema)

15.15 – 15.40 **Presentation** *Joost Kruis, University of Amsterdam (and ACT-Next)*
A general framework for choice dynamics
Discussant: Michèle Nuijten

15.40 – 16.30 **Invited speaker:** *Timo Bechger, senior researcher at Cito*
Sense and non-sense of item response theory

16.30 – 16.50 **Plenary meeting IOPS staff and students**

16.50 – 18.00 **Poster Session and Drinks** (Grand Café Aan de Beek)
Erik-Jan van Kesteren, Utrecht University
Qianrao Fu, Utrecht University
Esther Maassen, Tilburg University
Bunga Citra Pratiwi, Leiden University
Giulio Flore, Leiden University
Aline Claesen, Katholieke Universiteit Leuven
Shuai Yuan, Tilburg University
Richard Artner, Katholieke Universiteit Leuven

18.30 Conference dinner (Grand Café Aan de Beek)

Program Friday December 14th (Watercinema)

09.00 – 09.30 Registration / Coffee

09.30 – 10.15 **Presentation IOPS Best Paper Award Winner 2018**
Jed Cabrieto – University of Leuven

10.15 – 10.40 **Presentation** *Monika Vaheoja, University of Twente (and 10voordeleraar)*
Resetting performance standard in small samples with IRT and Circle-arc.
Discussant: Tom Wilderjans

10.40 – 11.05 Break (Watercinema)

11.05 – 11.30 **Presentation** *Fayette Klaassen, Utrecht University*
The Bayesian world of Probabilities, Odds and Updating.
Discussant: Sanneke Schouwstra

11.30 – 11.55 **Presentation** *Kimberley Lek, Utrecht University*
The optimal role of the EPST-result and teacher advice in the transition from primary to secondary education
Discussant: Rob Meijer

11:55 – 12.20 **IOPS Best Poster/Presentation Award Ceremony 2018**

12.20 – 12.30 **Closing by Cor Sluijter**

12.30 Take away Lunch (Grand Café Aan de Beek)

Thursday December 13th

13.05 – 13:30 A comparison of agreement coefficients for categorical and interval scales
Alexandra de Raadt – University of Groningen

Agreement assessment is of concern for both categorical as well as interval ratings. Kappa coefficients are commonly used for assessing agreement on a categorical scale, whereas correlation coefficients are commonly applied to assess agreement on an interval scale. In this study we compared the values of different agreement coefficients for both categorical and interval ratings using several real-world data sets. We studied empirical similarities between the various ways of measuring agreement. In addition, we studied how often we may reach similar decisions with different coefficients with regard to agreement assessment. Many authors have criticized the use of weighted kappa, a popular coefficient for ordinal ratings. We discussed the pros and cons of the use of quadratic kappa and the Pearson correlation. We can imagine that the much-criticized weighted kappa coefficient could generally be replaced by the Pearson correlation.

13.30 – 13.55 Comparing Constraint-based Causal Discovery algorithms in scenarios typical to psychology
Nitin Bhushan – University of Groningen

Researchers in psychology are often interested in understanding substantive causal relationships between variables underlying their phenomenon of interest. Such causal theories are of interest because they help predict the effects of interventions and are beneficial to both science and policy. One way of gaining insight into underlying mechanisms and effects of interventions is through true experiments (or randomized controlled trials; RCTs). However, in the context of certain branches of psychology, various real-world constraints do not permit use of RCTs and as a consequence, researchers often resort to observational studies.
When RCTs are not feasible and substantive theories yet to be developed, causal discovery algorithms can discover probabilistic causal relationships between variables of interest from observational data. In this talk, we assess three such procedures which use conditional independence as a constraint to infer underlying causal structures; the PC algorithm (Spirtes et al., 2000), LinGaM (Shimizu et al., 2006), and the FCI algorithm (Spirtes et al., 1995; Zhang, 2008). The PC algorithm assumes a linear model with Gaussian errors and no unmeasured common-causes or confounders. The LinGaM algorithm relaxes the Gaussian error assumption and retains assumptions of linearity and absence of hidden confounders. The FCI algorithm allows for hidden confounders while retaining linear Gaussian assumptions. To validate these procedures, we perform a simulation study varying the sample size, number of variables, degree of confounding, degree of non-normality of the error distribution, and graph sparsity. We score these procedures using two graph theoretic metrics (i) the structural Hamming distance and (ii) structural intervention distance. We discuss the results of our study and further discuss implications of such procedures for theory development in psychology.

13.55 – 14.20 Shrinkage priors for Bayesian measurement invariance: Practical and robust approaches for modeling and detecting non-invariance.
Sara van Erp – Tilburg University

When comparing multiple groups it is important to establish measurement invariance (MI), meaning that the latent construct under investigation is measured in the same way across groups. Traditionally, MI is tested using multiple group confirmatory factor analysis (MGCFA) with certain restrictions on the model. The goal is often to attain scalar invariance, which sets the loadings and intercepts equal across groups, so that factor means can be meaningfully compared. In practice, however, scalar invariance is often an unattainable ideal. Therefore, several alternative methods have been proposed to test for MI, such as partial MI, Bayesian approximate MI, and the alignment method. Although these techniques relax the restrictions imposed by the scalar invariance model, the assumptions they impose about the underlying structure of MI remain specific and stringent.
In this presentation, a novel method for modeling MI will be presented. The proposed method relies on the observation that MI essentially poses an identification problem, similar to the problem in sparse regression where the number of predictor variables is (much) greater than the number of observations. In sparse regression problems, regularization methods (e.g., the lasso) are popular approaches that identify the model by shrinking the small coefficients towards zero. We apply a similar strategy to the MI problem to model the invariance in a more flexible and robust manner than the current state-of-the-art methods.

14.20 – 14.45 Bayesian Covariance Structure Modelling of Responses and Process Data
Konrad Klotzke – University of Twente

A novel Bayesian modelling framework for response accuracy (RA), response times (RTs) and other process data is proposed. Nested (e.g., within a testlet) and crossed (e.g., between RAs and RTs for an item) local dependences are explicitly modelled in an additive covariance matrix. The inclusion of random effects (on person or group level) is not necessary, which allows constructing parsimonious models for responses and multiple types of process data. Bayesian Covariance Structure Models (BCSMs) are presented for various well-known dependence structures. In a simulation study, BCSMs are compared to state-of-the-art mixed-effect models. With an empirical example based on data from the Programme for the International Assessment of Adult Competencies (PIAAC) study, the flexibility and relevance of the BCSM for complex dependence structures in a real-world setting are discussed.

15.15 – 15.40 A general framework for choice dynamics
Joost Kruis – University of Amsterdam (and ACT-Next)

It has been demonstrated frequently that people often violate the rationality assumptions in decision making as implied by Luce's choice axiom. In this talk we present a simple framework for choices, which allows us to explain the occurrence of these violations. Inspired by the Ising model from statistical physics, we graphically represent a choice situation as a network, where the nodes correspond to cues and alternatives, and the edges between nodes describe the relationship between these. By introducing a Markov choice process that has rational choice behaviour as it's invariant distribution, and enforcing the rule that the decision process stops the first time the choice conditions are met, we obtain choice behaviour that is consistent with the research showing deviations of rationality.

15.40 – 16.30 Sense and non-sense of item response theory
Timo Bechger – senior researcher at Cito

Item response theory (IRT) came in the 1960s and caused a revolution in educational measurement. It alleviated psychometricians from the need to collect complete data and led to cool applications such as computer adaptive testing, student monitoring systems and international educational surveys. IRT has since become the dominant paradigm for educational measurement. As standardized testing became more popular in schools and computers became faster, the applications got bigger. Theoretical developments, on the other hand, were scant. One could say that psychometricians have only one, rather old, tool that they use for ever more complex applications. In this talk, I will illustrate two consequences of this. First, that IRT may be unsuited as a tool for some applications. Much like a hammer is not an ideal tool to build a skyscraper. Second, that some rather urgent issues are not addressed or ignored; simply because IRT cannot handle them. Most notably learning and change.

Friday December 14th

10:15 – 10:40 Resetting performance standard in small samples with IRT and Circle-arc
Monika Vaheoja – University of Twente/10voordeleraar

Resetting performance standard in exams with few respondents is statistically challenging because the estimates often include bias. Therefore do experts such as in Angoff method (1971) often reset the standards, and empirical information is often neglected. However, the standard-setting methods with experts are biased too and often expensive (Cizek & Bunch, 2007). In this presentation, we will compare Circle-arc equating (specially developed for small samples; Livingston & Kim, 2011) and IRT concurrent calibration with OPLM in resetting the cut-score from reference test to a new test form in different contexts. Responses are simulated in three different situations: sample size, test length, test difficulty and ability. The results demonstrate that even in small samples (50 subjects taking both tests) IRT-method outperforms Classical test theory when tests' difficulty and population ability interact.

11:05 – 11:30 The Bayesian world of Probabilities, Odds and Updating
Fayette Klaassen – Utrecht University

A Bayes factor can be used to quantify the relative evidence for any two hypotheses, it can be updated sequentially, and can be used to compare more than two hypotheses. In my PhD I have researched both practical and philosophical considerations in using a Bayes factor. In this talk I give an overview of some of these questions and answers. For example, what do power and error probabilities mean in Bayesian hypothesis testing? How can knowledge about a set of hypotheses be updated? What is the role of prior probabilities and how can they be specified? Three central concepts that are discussed in this talk are: (un)conditional error probabilities; prior/posterior odds; Bayesian updating.

11.30 - 11.55 The optimal role of the EPST-result and teacher advice in the transition from primary to secondary education
Kimberley Lek – Utrecht University

To determine the level of secondary education a pupil should transition to at the end of primary school, in the Netherlands two sources of information are consulted: 1) the result of an end-of-primary-school-test (EPST) and 2) the advice of the pupil's teacher. Depending on national policy decisions, one of these two sources is leading. Since 2015, the EPST-result is subordinated to the advice of the teacher, to great discontent of many psychometricians who warned for the subjectivity of teacher advice and teachers' sensitivity to pressure from parents and irrelevant child characteristics such as ethnicity. In my PhD, I investigate whether these psychometricians are right: has the change in policy in 2015 indeed led to worse transition advice? Or is there some merit in looking at the teacher advice? Additionally, I investigate whether instead of choosing between teacher and test it is possible to optimally weight and combine the EPST-result and teacher advice.

Cito
Amsterdamseweg 13
6814 CM Arnhem
Postbus 1034
6801 MG Arnhem
T (026) 352 11 11